Using HPE MPT Environment Variables for Pinning

For MPI codes built with HPE's MPT libraries, one way to control pinning is to set certain MPT memory placement environment variables. For an introduction to pinning at NAS, see Process/Thread Pinning Overview.

MPT Environment Variables

Here are the MPT memory placement environment variables:

MPI_DSM_VERBOSE

Directs MPI to display a synopsis of the NUMA and host placement options being used at run time to the standard error file.

Default: Not enabled

The setting of this environment variable is ignored if MPI_DSM_OFF is also set.

MPI_DSM_DISTRIBUTE

Activates NUMA job placement mode. This mode ensures that each MPI process gets a unique CPU and physical memory on the node with which that CPU is associated. Currently, the CPUs are chosen by simply starting at relative CPU 0 and incrementing until all MPI processes have been forked.

Default: Enabled

WARNING: If the nodes used by your job are not fully populated with MPI processes, use **MPI_DSM_CPULIST**, **dplace**, or **omplace** for pinning instead of **MPI_DSM_DISTRIBUTE**.
The **MPI_DSM_DISTRIBUTE** setting is ignored if **MPI_DSM_CPULIST** is also set, or if **dplace** or **omplace** are used.

MPI_DSM_CPULIST

Specifies a list of CPUs on which to run an MPI application, excluding the shepherd process(es) and **mpirun**. The number of CPUs specified should equal the number of MPI processes (excluding the shepherd process) that will be used.

Syntax and examples for the list:

- Use a comma and/or hyphen to provide a delineated list:

```
# place MPI processes ranks 0-2 on CPUs 2-4
# and ranks 3-5 on CPUs 6-8
setenv MPI_DSM_CPULIST "2-4,6-8"
```

- Use a "/" and a stride length to specify CPU striding:

```
# Place the MPI ranks 0 through 3 stridden
# on CPUs 8, 10, 12, and 14
setenv MPI_DSM_CPULIST 8-15/2
```

- Use a colon to separate CPU lists of multiple hosts:

```
# Place the MPI processes 0 through 7 on the first host
# on CPUs 8 through 15. Place MPI processes 8 through 15
# on CPUs 16 to 23 on the second host.
setenv MPI_DSM_CPULIST 8-15:16-23
```

- Use a colon followed by **`allhosts`** to indicate that the prior list pattern applies to all subsequent hosts/executables:

```
# Place the MPI processes onto CPUs 0, 2, 4, 6 on all hosts
setenv MPI_DSM_CPULIST 0-7/2:allhosts
```

Examples

An MPI job requesting 2 nodes on Pleiades and running 4 MPI processes per node will get the following placements, depending on the environment variables set:

```
#PBS -lselect=2:ncpus=8:mpiprocs=4
module load <mpt_module>
setenv ....
cd $PBS_O_WORKDIR
mpiexec -np 8 ./a.out
```

- setenv MPI_DSM_VERBOSE
  setenv MPI_DSM_DISTRIBUTE

```
MPI: DSM information
MPI: MPI_DSM_DISTRIBUTE enabled
grank   lrank   pinning  node name      cpuid
    0       0   yes      r86i3n5            0
    1       1   yes      r86i3n5            1
    2       2   yes      r86i3n5            2
    3       3   yes      r86i3n5            3
    4       0   yes      r86i3n6            0
    5       1   yes      r86i3n6            1
    6       2   yes      r86i3n6            2
    7       3   yes      r86i3n6            3
```

- setenv MPI_DSM_VERBOSE
  setenv MPI_DSM_CPULIST 0,2,4,6

```
MPI: WARNING MPI_DSM_CPULIST CPU placement spec list is too short.
MPI:         MPI processes on host #1 and later will not be pinned.
MPI: DSM information
grank   lrank   pinning  node name      cpuid
    0       0   yes      r22i1n7            0
    1       1   yes      r22i1n7            2
    2       2   yes      r22i1n7            4
    3       3   yes      r22i1n7            6
    4       0   no      r22i1n8             0
    5       1   no      r22i1n8             0
    6       2   no      r22i1n8             0
    7       3   no      r22i1n8             0
```

- setenv MPI_DSM_VERBOSE
  setenv MPI_DSM_CPULIST 0,2,4,6:0,2,4,6

```
MPI: DSM information
grank   lrank   pinning  node name      cpuid
    0       0   yes      r13i2n12           0
    1       1   yes      r13i2n12           2
    2       2   yes      r13i2n12           4
    3       3   yes      r13i2n12           6
```

```
    4       0   yes      r13i3n7           0
    5       1   yes      r13i3n7           2
    6       2   yes      r13i3n7           4
    7       3   yes      r13i3n7           6
```

- setenv MPI_DSM_VERBOSE
  setenv MPI_DSM_CPULIST 0,2,4,6:allhosts

```
MPI: DSM information
grank   lrank   pinning  node name      cpuid
    0       0   yes      r13i2n12           0
    1       1   yes      r13i2n12           2
    2       2   yes      r13i2n12           4
    3       3   yes      r13i2n12           6
    4       0   yes      r13i3n7            0
    5       1   yes      r13i3n7            2
    6       2   yes      r13i3n7            4
    7       3   yes      r13i3n7            6
```

Article ID: 286
Last updated: 30 Sep, 2021
Revision: 44
Porting/Building Code -> Optimizing/Troubleshooting -> Process/Thread Pinning -> Using HPE MPT Environment Variables for Pinning
https://www.nas.nasa.gov/hecc/support/kb/entry/286/